Intelligent Systems with Applications 17 (2023) 200164

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Leveraging hybrid machine learning and data fusion for accurate mapping of malaria cases using meteorological variables in western India

Abhilash Singh [a,*], Manish Mehra [a], Amit Kumar [a], M Niranjannaik [a], Dev Priya [b], Kumar Gaurav [a,*]

[a] Fluvial Geomorphology and Remote Sensing Laboratory, Department of Earth and Environmental Sciences, Indian Institute of Science Education and Research Bhopal, Madhya Pradesh 462066, India

[b] Patent Facilitation Centre, Technology Information Forecasting and Assessment Council, New Delhi, India [1]

ARTICLE INFO

Keywords:
Machine learning
Data fusion
Disease prediction
Meteorological variables
Intelligent system

ABSTRACT

We propose a hybrid machine learning algorithm (*i.e.,* $P^2CA - PSO - ANN$) to model malaria outbreak in three districts (Barmer, Bikaner, and Jodhpur) of Rajasthan in the Western India. We have used different meteorological variables (*i.e.,* relative humidity, temperature, and rainfall) as input features to predict malaria. We have also considered the combined impact of these variables through a linear data fusion. We then extract the uncorrelated information from the feature set by applying Probabilistic Principal Component Analysis (P^2CA). We trained the fully connected feed-forward Artificial Neural Network (ANN) by optimising its hyperparameters iteratively through a bio-inspired optimisation algorithm (Particle Swarm Optimisation). We train and evaluate the performance of this algorithm using monthly meteorological variables from 2009 - 2012. This accurately predicts the malaria cases with the coefficient of correlation ($R = 0.99$), and Root Mean Square Error ($RMSE = 1.76$). Finally, we compare our model with different benchmark algorithms (Generalised Regression Neural Network (GRNN), Gaussian Process Regression (GPR), Support Vector Regression (SVR), Random Forest, and Radial Basis Neural Networks (RBNN)) in terms of accuracy. We observed the performance of hybrid machine learning model relatively high. This study can be used as an early warning intelligent system to predict the malaria outbreaks solely from meteorological data.

1. Introduction

The linkage between climatic variables and transmission (or spread) of vector-borne diseases, for example, malaria, dengue fever, lyme, and scrub typhus is well established. These diseases are either season-specific or erupt because of extreme events such as flood, drought etc. (Patz, 2002). Short-term changes in the climatic variables because of global warming has intensified the need of studying climate change and disease transmission concurrently (Rocklöv & Dubrow, 2020).

Among the various vector-borne diseases, malaria has caused a significant health burden globally (Caminade et al., 2014). It is a mosquito-borne disease caused by the different species of the *Plasmodium* protozoan parasites, namely *P. falciparum, P. vivax, P. malariae, P. knowlesi, P. ovale wallikeri,* and *P. ovale curtisi*. *P. falciparum and P. vivax* cover a larger portion of the cases (≈ 95%) in the world (Garrido-Cardenas, González-Cerón, et al., 2019, Garrido-Cardenas, Cebrián-Carmona, et al., 2019). It is among the top ten causes for death in lower-income countries. According to the World Health Organization (WHO), from 2000 to 2019, about 1.5 billion malaria cases and 7.6 million malaria deaths were reported. There were approximately 229 million malaria cases in 2019 in about 87 countries (WHO, 2020). Globally, only 29 countries contribute nearly 95% of the total malaria cases in the world. Among these, Nigeria contributes the highest percentage (≈ 27%), the Democratic Republic of the Congo (≈ 12%), Uganda (≈ 5%), Mozambique (≈ 4%) and Niger (≈ 3%). The South-East Asia Region (SEAR) contributes about 3% to the global malaria cases of which India alone contributes about (≈ 60%) followed by Indonesia (30%) and Myanmar (10%). India has reported a reduction in malaria cases from 20 million in 2000 to nearly 5.6 million in 2019 (WHO, 2020). Despite this malaria is still a major healthcare challenge in India.

* Corresponding authors.
E-mail addresses: sabhilash@iiserb.ac.in, abhilash.iiserb@gmail.com (A. Singh), manish16@iiserb.ac.in (M. Mehra), amit17@iiserb.ac.in (A. Kumar), niranjannaik@iiserb.ac.in (M Niranjannaik), devpri6@gmail.com (D. Priya), kgaurav@iiserb.ac.in (K. Gaurav).
[1] Now at CIP LEGIT, Gurugram Haryana, India.

https://doi.org/10.1016/j.iswa.2022.200164
Received 24 February 2022; Received in revised form 23 June 2022; Accepted 1 December 2022
Available online 5 December 2022

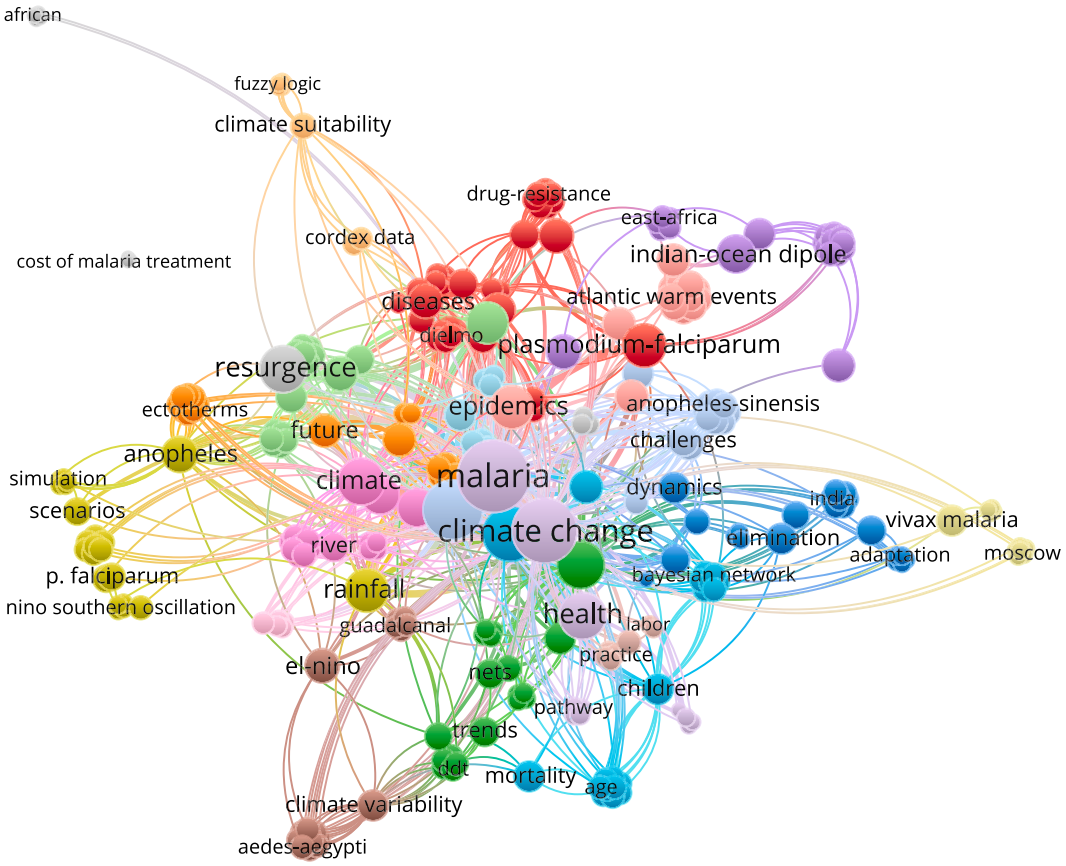

Fig. 1. Co-keywords burst based bibliometric analysis of the keywords "*{malaria}* **AND** *{Climate}*". Total 2117 research publications published during 1991 to 2022 (till February 10, 2022) in WoS database.

In India, malaria is more prevalent in north-eastern states, *i.e.*, Orissa, Chhattisgarh, and Jharkhand. After the eradication attempts in the late 1970s, malaria has been reported in the arid regions of Gujarat and Rajasthan (Tyagi et al., 1995, Akhtar & McMichael, 1996). In Rajasthan, the major epidemics have occurred in the desert regions such as; Barmer, Bikaner, Jaisalmer, Jodhpur, Pali, and Sri Ganganagar.

Persistence of malaria in India is because of its huge geographic and climatic variability, which provides suitable ecological conditions for many parasites (Sarkar et al., 2019). Climatic factors enhance the breeding sites for mosquitoes (Haque et al., 2010). Mosquitoes morphological processes (*e.g.,* growth) are highly dependent on ambient temperature, humidity, and stagnant water bodies. Rainfall has received a lot of attention as a key factor in rising mosquito breeding sites. The breeding areas can be limited by drought or severe flooding in the region (Caminade et al., 2019, Kelly-Hope et al., 2009, Alonso et al., 2011). The semi-arid and arid regions of Western India are considered as an unpredictable malaria zones with low and high incidences mainly affected by the rainfall (Mathur et al., 1992). As a result, malaria outbreaks caused due to climatic conditions may results disease transmission and eventually complicate the situation (Jetten et al., 1996, Hulme et al., 1996, Sutherst, 1998).

Owing to the lack of medical facilities, especially during the time of pandemic, the symptoms of malaria are much more severe (Di Gennaro et al., 2020). Preventing or reducing the risk factor for malaria is extremely difficult, particularly in lower-income countries. The technology can provide alternative solutions by allowing for early warning mechanisms to monitor the spread of disease and advance management of treatment facilities to ensure a more timely health services that can save lives. The availability of any predictive model will not only help healthcare services but also to avoid or reduce the large-scale spread of diseases (Modu et al., 2017).

This study proposes a hybrid machine learning algorithm to predict the malaria cases using meteorological variables. We selected temperature, rainfall, and relative humidity as the potential input features. To study the combined effect of these meteorological variables, we generated some additional features using linear data fusion of two features. Finally, we train and evaluate the performance of the machine learning model to accurately predict the malaria cases. To the best of our knowledge, no such studies have been conducted to assess the combined effect of relative humidity with temperature, temperature with rainfall, relative humidity with rainfall, along with the individual climate variables on the malaria cases using a hybrid machine learning approach.

2. Related work

Machine learning models have been setup using the meteorological variables to accurately predict malaria cases. Fig. 1 shows the bibliometric analysis using the keywords malaria and climate. The number of publications has increased drastically in the last two decades, with only 21 publications in 2001 to 152 publications in 2021. In total, we found 2117 research publications from 1991 to 2022 in the Web of Science (WoS) database. Researchers have traditionally used linear regression and time series approaches (Srimath-Tirumula-Peddinti et al., 2015, Jones et al., 2007, Kumar et al., 2020).

Modu et al. (2017) have used the maximum and minimum temperature, precipitation, relative humidity, solar radiation, and wind speed as the potential climatic variables for predicting malaria outbreaks. They reported that the temperature, and relative humidity are positively correlated (Pearson's cross-correlation) to the number of malaria cases. Modu et al. (2017) compared and evaluated the performance of seven regression-based machine learning algorithms; linear regression, logistic regression, decision tree, support vector machine, optimised Support

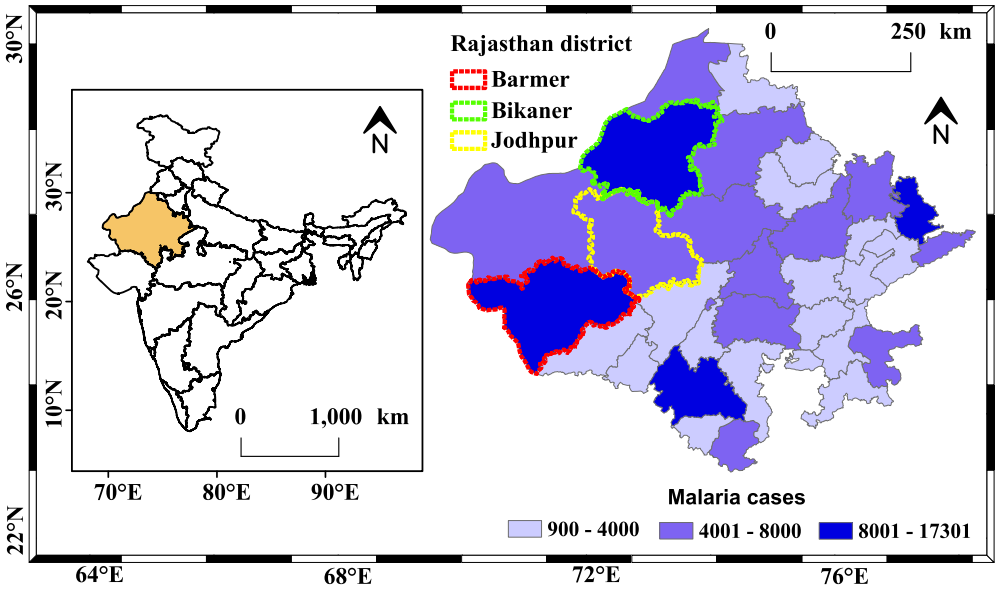

Fig. 2. Malaria prone districts (Barmer, Bikaner, and Jodhpur) of Rajasthan, India. Map on the right shows the accumulated malaria cases from 2009 to 2017.

vector machine, naive Bayes, K-nearest neighbours, and *k*-mean. They observed that optimised Support vector machine outperforms all the other machine learning algorithms through 10-fold cross-validation. Recently, Kim et al. (2019), proposed a weather-based malaria prediction model using weekly time-series temperature and precipitation data. They reported the model prediction accuracy (R > 0.8) is higher for short-term (1 or 2 week ahead). Thakur and Dharavath (2019) used the climatic variable along with the clinical data to predict the malaria cases. They used rainfall, relative humidity, temperature, and vegetation index as environmental variables and trained an ANN model for accurate mapping of malaria cases. They reported an error varying from 18% to 117%. More recently, Nkiruka et al. (2021) proposed a malaria incidence classification model and compared their results with different machine learning models. They considered three climatic variables such as precipitation, temperature, and surface radiation for mapping the number of malaria cases. Temperature has a strong linear relationship with malaria cases among all the three climatic variables. They have used *k*-mean learning to clean and remove the outliers and XGBoot ensemble learning approach for classification. They reported that the association between the climatic variables and malaria cases varies from one geographic location to another.

All the models discussed above only consider the individual effect of climatic variables. This study aims to strengthen the prediction accuracy of previous studies by extracting the maximum information from the climatic variables using a hybrid machine learning algorithm.

3. Study area

We have selected three districts, Bikaner, Barmer, and Jodhpur of Rajasthan province in the northwestern India to predict the malaria outbreak (Fig. 2). These districts are chosen based on the high number of malaria cases and data availability. Tyagi et al. (1995) reported that after the construction of three major canal systems, the Gang, the Bhakra Sirhind feeder canal, and the Indira Gandhi canal have provided a favourable ecology for malaria breeding.

Malaria outbreak in the study area usually occurs during the Indian Summer Monsoon period (June-September) (Lingala et al., 2020, Kumar et al., 2022, Parihar et al., 2022). During the monsoon period, rainfall, temperature, relative humidity, and waterlogging provide a favourable condition for the parasite growth in mosquito (Arab et al., 2014). Fig. 3 show the time series of malaria outbreak in Bikaner, Barmer, and Jodhpur from 2019 - 2014.

Fig. 3 shows a cyclic behaviour of malaria outbreak in all three districts. The disease starts to spread about a month after the onset of monsoon in June. It reaches to its peak in August and September and then starts to decrease.

The annual average minimum and maximum temperature in the study area varies between 23° to 40 °C. The annual average rainfall varies between 313 mm to 675 mm for the western and eastern part of Rajasthan, respectively. The average annual relative humidity varies from 45% to 50%.

4. Material and methods

4.1. Data

We obtained the monthly malaria cases of *P. vivax* and *P. falciparum* of three districts (Barmer, Bikaner, Jodhpur) for a period between January 2009 to December 2012 from National Vector Borne Disease Control Programme (NVBDCP), New Delhi (Lingala, 2017). The corresponding monthly meteorological data; temperature, rainfall and relative humidity (at 8:30 IST and 5:30 IST) is downloaded from the Indian Meteorological Department (IMD) (https://mausam.imd.gov.in/).

4.2. Features processing

The performance of any machine learning model depends on feature pre-processing (Hall et al., 1971). It is highly desirable to adopt essential pre-processing steps (specially for numerical features) to develop a efficient and robust machine learning model (Alshdaifat et al., 2021).

4.2.1. Outliers technique

We used Median Absolute Deviation (MAD) method to identify and remove the outliers present in the data (Fig. 4). We estimated the median of absolute deviation from the median and finally the MAD is calculated by multiplying it by an empirically derived constant (Leys et al., 2013).

$$MAD = B \cdot \boldsymbol{M}\left(\left|A - \boldsymbol{M}(A)\right|\right) \tag{1}$$

where M is the median of the series A consisting n observation and B is an empirical constant whose value is derived from

$$B = \frac{-1}{(\sqrt{2} \cdot erfcinv(3/2))} \tag{2}$$

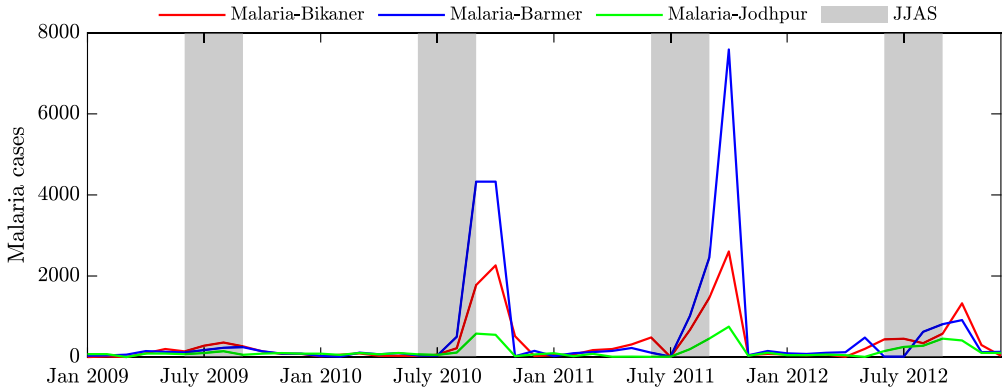

Fig. 3. Time series (form 2009 to 2012) of the malaria cases in three districts (Bikaner, Barmer, and Jodhpur) of Rajasthan in the Western India. The shaded region in grey colour represents the monsoon period (*i.e.*, JJAS; June-July-August-September).

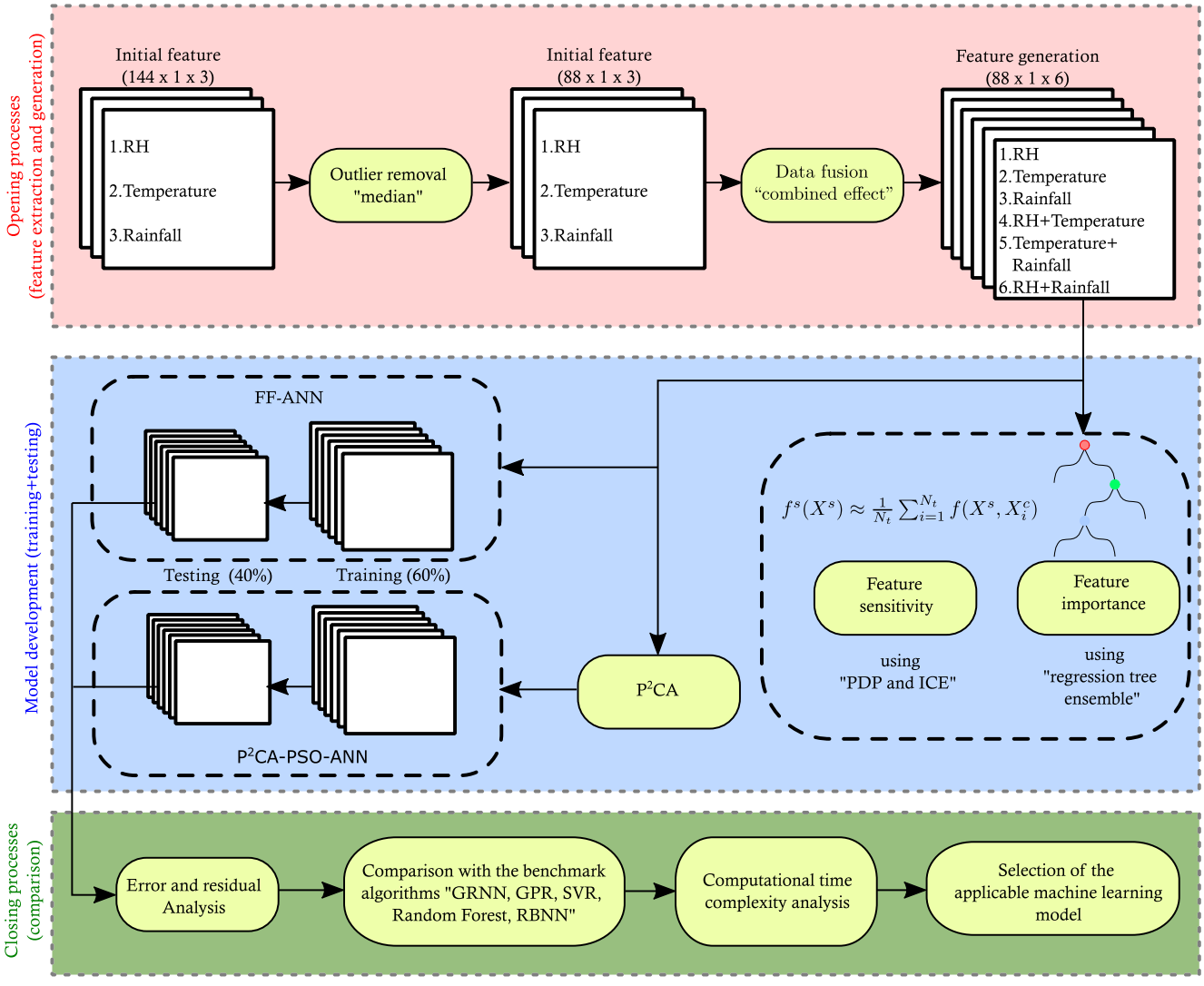

Fig. 4. Detailed workflow illustrates the procedure of input features selection, machine learning models, and result analysis.

where $erfcinv$ represents the inverse complementary error function. Finally, the outliers are identified by using the following criterion

$$\left(M - 3 \cdot MAD\right) < A < \left(M + 3 \cdot MAD\right) \tag{3}$$

Any values that lies outside this range are marked as outlier.

4.2.2. Feature generation and data fusion

Initially, we select three meteorological variables; temperature, rainfall and relative humidity as the potential features to train the machine learning model. These features will only evaluate the effect of individual measures. To study the effect of combined measures, we have created three additional features through linear data fusion technique, such as

relative humidity with temperature, temperature with rainfall, and relative humidity with rainfall (Fig. 4).

4.2.3. Feature importance and correlation

The goodness of a machine learning model depends on the relevancy of features from which it is trained on. The relevancy of any feature can be assessed by estimating the feature importance score. Higher the feature importance more relevant is the feature. We used the regression tree ensemble technique to estimate the feature importance (Singh et al., 2022a, 2022b). We used the Least-Squares Boosting (LSBoost) algorithm to train a regression ensemble. In doing so, we created the regression ensemble by boosting hundred regression trees. This step is based on an assumption that the regression tree is a weak learner (with unity learning rate). Further, we estimated the feature importance of each feature in a tree by summing all the changes in the node risk occurred because of splits on every feature. The final estimate is obtained by diving the changes with the number of branch nodes, N_{branch}. The node risk change for the parent node is calculated by subtracting the total risk of the two children ($R_{c1} + R_{c2}$) from the parent risk (R_p) given as;

$$\Delta R = \frac{R_p - (R_{c1} + R_{c2})}{N_{branch}} \tag{4}$$

The risk at individual node (R_i) is calculated according to

$$R_i = P_i \cdot MSE_i \tag{5}$$

where P_i represents the node probability and MSE_i represents the mean square error of the node i.

Apart from feature importance, we also estimated the feature association matrix to identify any correlated feature. Presence of any highly correlated features makes the machine learning model highly unstable and sensitive (Tolo̧şi & Lengauer, 2011, Singh, Gaurav, et al., 2021).

4.2.4. Feature sensitivity

Feature importance graph tell us about the relative importance of each feature. To identify whether these features have negative or positive impact on the machine learning model, we performed the sensitivity analysis of all the features using Partial Dependency Plot (PDP) (Friedman, 2001, Singh, Nagar, et al., 2021) and Individual Conditional Expectation (ICE) Curve (Goldstein et al., 2015, Singh et al., 2020).

PDP evaluates the partial dependence of predictand (*i.e.,* malaria cases) on a single feature by marginalising the effect of all other features. Let $\boldsymbol{X^s}$ be a singleton represented by $\boldsymbol{X^s} = \{x_{s1}\}$ of the whole feature set, $\boldsymbol{X}$ represented by $\boldsymbol{X} = \{x_1, x_2, \cdots, x_m\}$. Consider $\boldsymbol{X^c}$ be the complementary set of $\boldsymbol{X^s}$ in the feature set $\boldsymbol{X}$. The predictand response, $f(\boldsymbol{X})$, depends on all the elements in the feature set $\boldsymbol{X}$ according to;

$$f(\boldsymbol{X}) = f(\boldsymbol{X^s}, \boldsymbol{X^c}) \tag{6}$$

The partial dependence of the predictand on $\boldsymbol{X^s}$ is calculated by the expectation of the predictand response with respect to $\boldsymbol{X^c}$.

$$f^s(\boldsymbol{X^s}) = E_c[f(\boldsymbol{X^s}, \boldsymbol{X^c})] \tag{7}$$

$$= \int f(\boldsymbol{X^s}, \boldsymbol{X^c}) \cdot p_c(\boldsymbol{X^c}) \cdot d\boldsymbol{X^c} \tag{8}$$

where $p_c(\boldsymbol{X^c})$ represent the marginal probability of $\boldsymbol{X^c}$ given by

$$p_c(\boldsymbol{X^c}) \approx \int p(\boldsymbol{X^s}, \boldsymbol{X^c}) \cdot d\boldsymbol{X^s} \tag{9}$$

The final partial dependence estimate (*i.e.,* the average marginal effect) for $\boldsymbol{X^s}$ is given by

$$f^s(\boldsymbol{X^s}) \approx \frac{1}{N_t} \sum_{i=1}^{N_t} f(\boldsymbol{X^s}, \boldsymbol{X_i^c}) \tag{10}$$

where N_t is the total number of observations and $\boldsymbol{X_i} = (\boldsymbol{X_i^s}, \boldsymbol{X_i^c})$ represents the i^{th} observation. The ICE is calculated by dis-aggregating the average effect of Equation (10) according to

$$f_i^s(\boldsymbol{X^s}) = f(\boldsymbol{X^s}, \boldsymbol{X_i^c}) \tag{11}$$

Finally the level effect is removed by

$$f_{i,centered}^s(\boldsymbol{X^s}) = f(\boldsymbol{X^s}, \boldsymbol{X_i^c}) - f(min(\boldsymbol{X^s}), \boldsymbol{X_i^c}) \tag{12}$$

This is done for better visualisation of the cumulative effect of $\boldsymbol{X^s}$. Generally, ICE is used for analysing the presence of any heterogeneity at any individual observation that had been obscure by the averaging effect of PDP.

4.3. Artificial neural network model

ANN is based on the concept and functionality of biological neuron present in human brain. The fundamental unit of ANN is artificial neuron which is a mathematical model that mimics the behaviour of biological neuron. Information is passed into the artificial neurons and it is processes using mathematical function to generate the final output (Asteris et al., 2017). To exactly mimic the random behaviour of biological neurons, the information is multiplied with a weight value before passing it to the artificial neuron. Several artificial neurons are grouped together to form ANN. Generally for setting up an ANN model, we need to define three things; (i) architecture of the network, (ii) mathematical function that describe the models, and (iii) training algorithm. We have discussed these in the succeeding subsections.

4.3.1. Feed forward artificial neural network (FF-ANN)

We proposed an architecture of a 6-20-1 fully connected FF-ANN (Fig. 5). In this type of structure, there is no feedback (*i.e.,* loop). The information flows only in one direction *i.e.,* from input towards the output. The neurons present in the same layers are not connected to each other, but they are connected with the neurons present in the previous and the upcoming layers. The fully connected feed forward ANN architecture for the prediction of malaria cases consists of single hidden layer with twenty neurons (Fig. 5).

4.3.2. Activation function at each layer

The selection of activation function at each layer strongly affects the model output (Karlik & Olgac, 2011). Generally, non-linear transfer functions are used in hidden layer. In this study, we have used the hyperbolic tangent sigmoid transfer function (Vogl et al., 1988) at the output of hidden and output layer as depicted in Fig. 5. Mathematically it is expressed as;

$$f(n) = tansig(n) = \frac{2}{\left(1 + e^{-2 \cdot n}\right) - 1} \tag{13}$$

This function is analogous to hyperbolic tangent function, it only differs in terms of computational time complexity. The execution time of *f(n)* is faster than the hyperbolic tangent function with very little variation in the numerical output. This is a trade-off for the feed-forward ANN, where speed is the primary interest than the exact shape of the transfer function (Dorofki et al., 2012). We use a linear (or identity) activation function at the output of the input layer.

4.3.3. Training algorithm

Various training algorithm such as Levenberg-Marquadt backpropagation, scaled conjugated gradient backpropagation, and Bayesian regularization backpropagation to optimise a multivariate function exist (Corte-Valiente et al., 2017). However, none of these can guarantee global optimal solution. For mapping number of malaria cases, we found that Levenberg-Marquadt backpropagation technique provides more promising results as compared to other algorithms. It is an it-

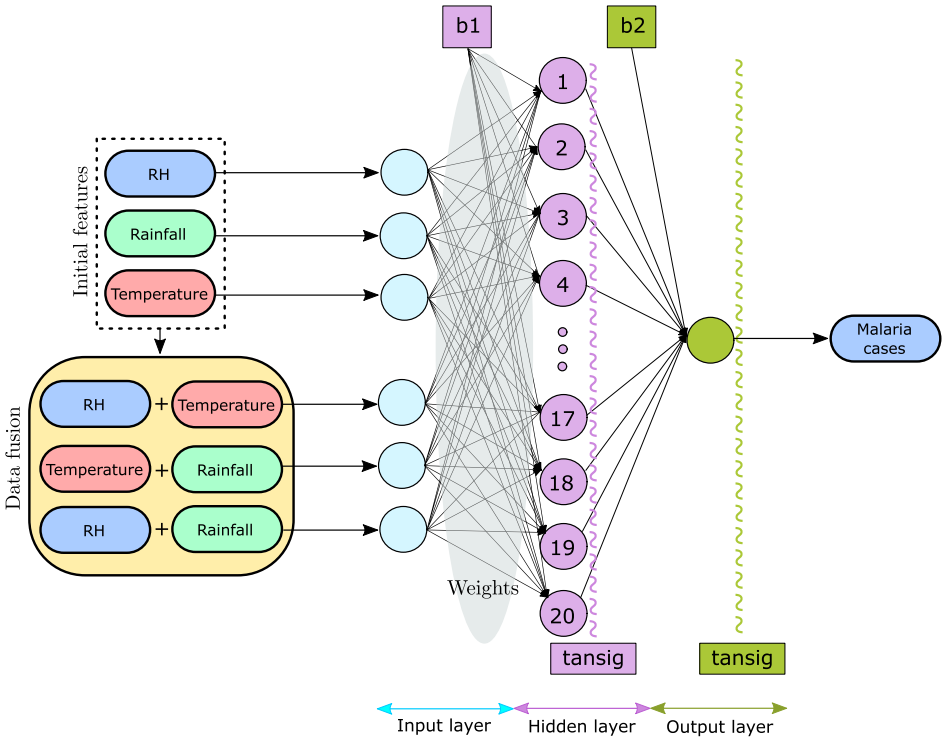

Fig. 5. Architecture of 6-20-1 backward propagation based fully connected FF-ANN. It consists of six inputs, twenty neurons in the hidden layer and one output.

erative algorithm that computes the optimal minima of a multivariate function to update the weight and bias values (Equation (14)).

$$w_{k+1} = w_k - [\boldsymbol{J}^T\boldsymbol{J} + \mu\boldsymbol{I}]^{-1}\boldsymbol{J}^T\boldsymbol{e} \tag{14}$$

where $\boldsymbol{J}$ represents the Jacobian matrix and μ is a scalar coefficient. It contains the first derivatives of the network errors with respect to the weights and biases. $\boldsymbol{e}$ represents the vector of network errors. Finally, we randomly divided the data into two parts in a 60:40 ratio for training, and testing of the FF-ANN 6-20-1 architecture, respectively (Fig. 4).

4.4. Hybrid model

We proposed a novel hybrid algorithm based on the coupling of Probabilistic Principal Component Analysis (P^{2}CA), Particle Swarm Optimisation (PSO), and ANN, such as P^{2}CA – PSO – ANN to predict the malaria cases using meteorological variables. In the succeeding subsections, we have discussed complete coupling process (Fig. 6).

4.4.1. Probabilistic principal component analysis

We applied P^{2}CA as a feature pre-processing step where the aim was to extract the most uncorrelated information from the input feature set. P^{2}CA efficiently estimate the principal axis even if some or all the data vector consist of single or more missing values by using expectation-maximisation (EM) algorithm (Tipping & Bishop, 1999). We reconstructed the data by considering the first three Principal Components (PCs) of P^{2}CA, they consist of 95% of the variance.

4.4.2. Particle swarm optimisation

PSO algorithm is based on swarm intelligence that was proposed in 1995 by Kennedy and Eberhart (1995). It has fewer parameters and the complete optimisation process is governed by iterating formula which reduces the computation burden. It has very high efficacy in optimising various theoretical and practical problems (Zhang et al., 2018, Singh, Sharma, et al., 2021). In general it is composed of two equations for updating the position and velocity iteratively (Equations (15) and (16)).

$$v_{in}^{t+1} = v_{in}^{t} + c_1 \cdot r_1 \cdot (P_{best}^{t} - x_{in}^{t}) + c_2 \cdot r_2 \cdot (g_{best}^{t} - x_{in}^{t}) \tag{15}$$

Table 1
Simulation parameters of PSO for optimising weights and biases.

Parameter	Value
Swarm size	6
Maximum iteration (t_{max})	50
c_1	2
c_2	4-c_1
Fitness function	MSE

$$x_{in}^{t+1} = x_{in}^{t} + v_{in}^{t+1} \tag{16}$$

where P_{best} is the particle (or swarm) best solution, g_{best} is the global best solution, c_1 is the cognitive component, c_2 is the social component, r_1 and r_2 are the random number between 0 and 1, x_{in}^{t} is the current particle position, v_{in}^{t} is the current particle velocity, v_{in}^{t+1} is the velocity at the next iteration, and x_{in}^{t+1} is the position at next iteration.

As illustrated in Fig. 6b, each particle iterates the position and velocity information from its own best solution (P_{best}) to global best solution (g_{best}). O_{the}^{opt} is the targeted theoretical optima. After affected by various factors (particle memory and swarm influence), the velocity changes from v_{in}^{t} to v_{in}^{t+1} with a position change from x_{in}^{t} to x_{in}^{t+1}. It is worthy to mention that the particle memory and swarm influence lines are parallel the x_{in}^{t} to the P_{best} and g_{best}, respectively. The algorithm will keep iterating and updating the position and velocity until it reaches more closer to the theoretical optima (Fig. 6c).

4.4.3. P^{2}CA – PSO – ANN

To couple P^{2}CA, and PSO with ANN, we have considered 6-6-1-1 ANN architecture as illustrated in Fig. 6a. The first hidden layer consists of six neurons and the second hidden layer consists of single neuron. Both these layers are followed by a tangent sigmoid transfer function (*i.e.,* tansig). The input and the output layers use linear activation function (*i.e.,* purelin). The P^{2}CA reconstructed data are fed to the model input, and the weights and biases are optimised iterately by PSO. The simulation parameters of PSO are given in Table 1. Similar to the FF-

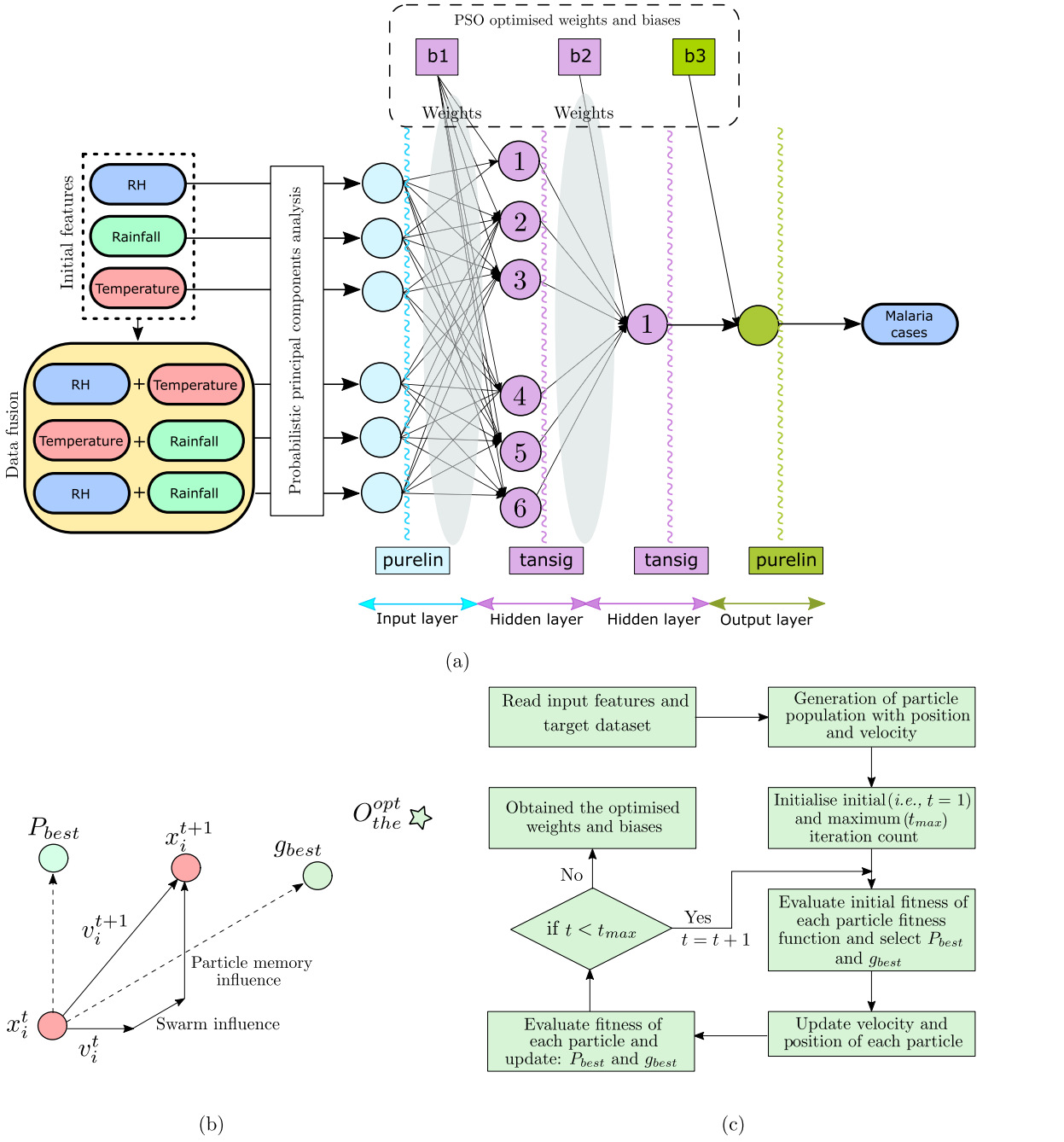

Fig. 6. (a) Schematic representation of the hybrid model. (b) Position and velocity updates in PSO. (c) Flowchart for optimising the weights and biases of the ANN.

ANN, we have randomly divided the data into two parts in a 60:40 ratio for training, and testing of the proposed hybrid algorithm, respectively.

5. Results and discussion

5.1. Feature importance, correlation, and sensitivity

We plot the relative feature importance score of each feature (Fig. 7a). We found relative humidity to be the single most important feature with the highest importance score amongst all the features. Higher the value of importance score, the more relevant is the feature in the prediction of malaria cases. It is followed by the relative importance score of the combined measure of relative humidity and temperature (*i.e.,* RH + Temperature). The measure of relative humidity and rainfall (*i.e.,* RH + Rainfall) has the least importance score. It is important to highlight that the relative importance of temperature is less than the combined features that includes temperature (*i.e.,* RH + Temperature and Temperature + Rainfall). This indicates when temperature variable is combined with other meteorological variables, it becomes more relevant. Further, we plot the feature association matrix (Fig. 7b). We observed no highly correlated features, this indicates model is less susceptible to instability.

We have plotted the Partial Dependency Plot (PDP) (shown in red line) and Individual Conditional Expectation (ICE) curves in Fig. 8. We observed no clear impact (or trend) of features with the malaria cases. Overall we found a fluctuating positive impact for RH + Temperature and Temperature + Rainfall and a fluctuating negative impact for RH and RH + Rainfall. We only observed a slight variation in the case of temperature.

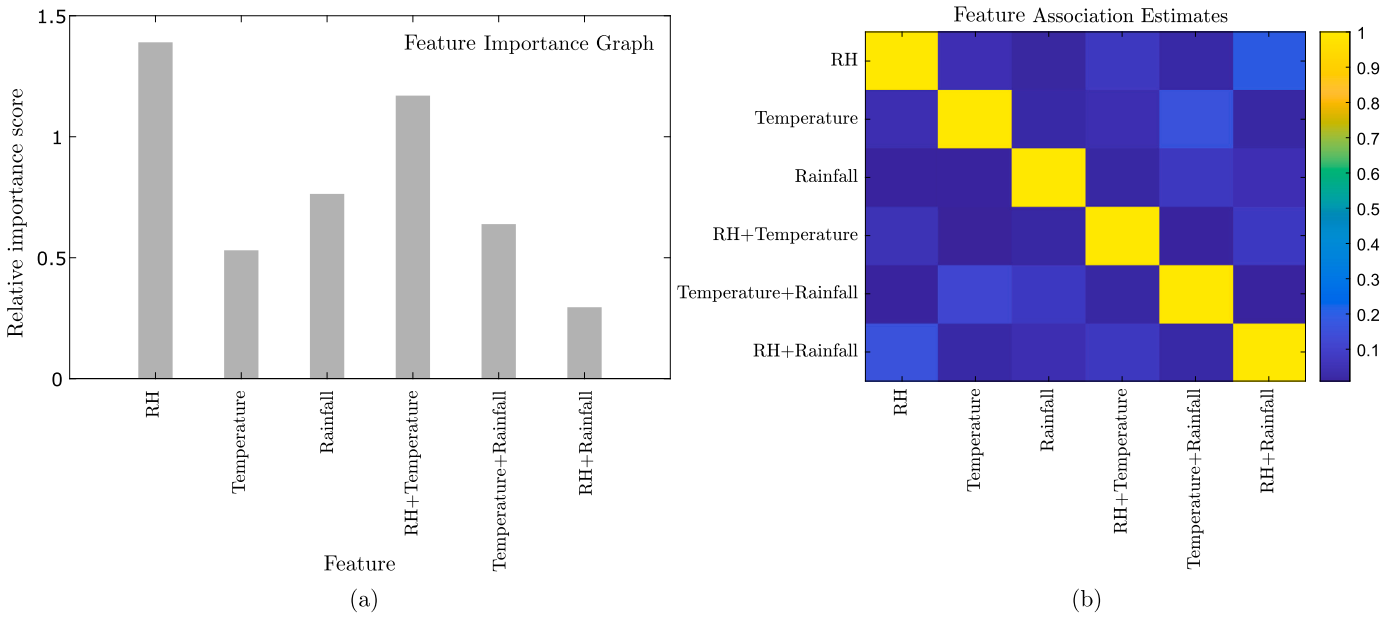

Fig. 7. (a) Bar graph shows the relative importance score of each feature, (b) shows the feature association matrix.

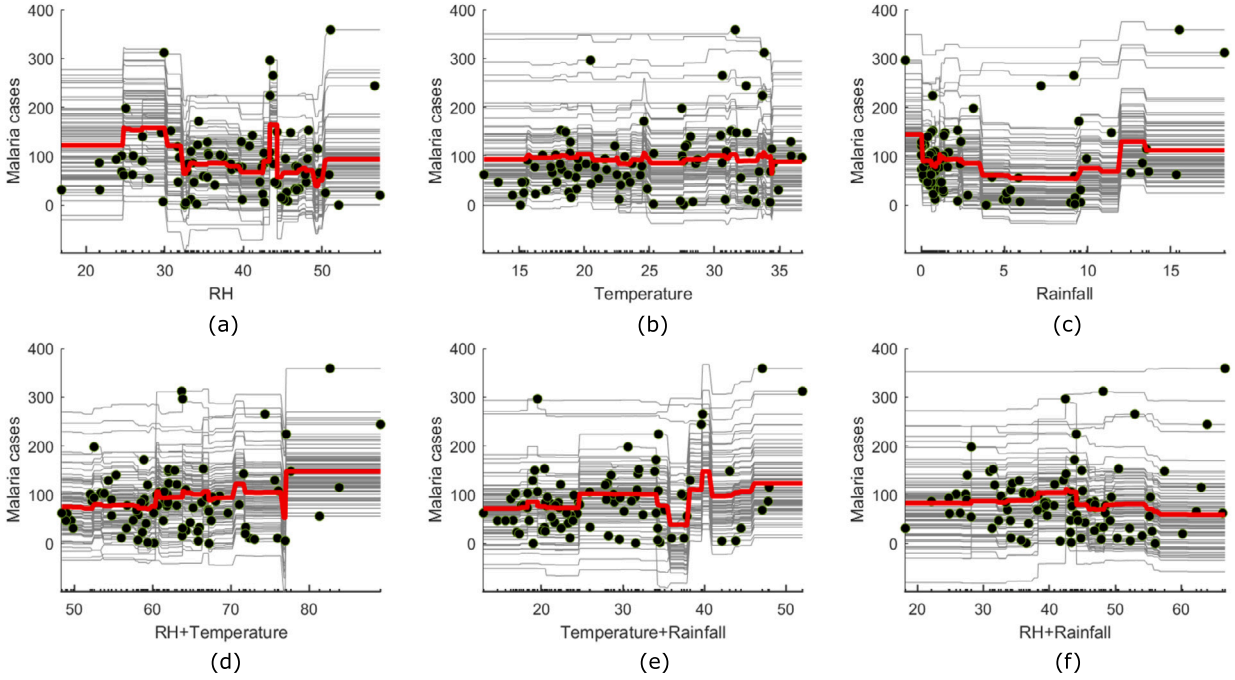

Fig. 8. Feature sensitivity analysis using PDP (in red line) and ICE curves (in grey lines).

5.2. Performance of the FF-ANN model

Once we trained the FF-ANN model using 60% of the data (N = 53), we evaluated the performance of the model on training data itself to report the training accuracy/statistics. We plot a linear regression curve between the estimated and observed values (Fig. 9a). To evaluate the performance of the model, we used R, RMSE, bias as the performance metrics. A detail of the performance metrics has been explained in Appendix A. The model performs reasonably well with R = 1, RMSE = 0, and zero bias. However, testing the model performance only on the training data is insufficient and result into bias. To evaluate its generalisation capability, we test the model performance on unseen data. We used the remaining 40% of the data (N = 35) for testing. We found that the trained model performs marginally on the test datasets with R = 0.72, RMSE = 62.23, and bias of -19.28 with moderate scattering (Fig. 9b). We found that few points lies outside the 95% confidence interval resulting in either overestimation or underestimation (marked in red circles).

To understand the errors and its impact on the performance of the FF-ANN model, we calculate the error from L1 norms by discarding the absolute part and plotted the error histogram with 10 bin size (Fig. 10). The shades of red and green correspond to the error associated with the training, and testing phase. Vertical line shown in orange represents the zero-error line. The total error ranges from -113.4 (left most bin) to 137.7 (right most bin). The negative sign indicates overestimation and the positive sign indicates underestimation. The training errors is more centric in nature and lies near the zero-error line followed by the testing error. The overall error follows a Gaussian distribution with

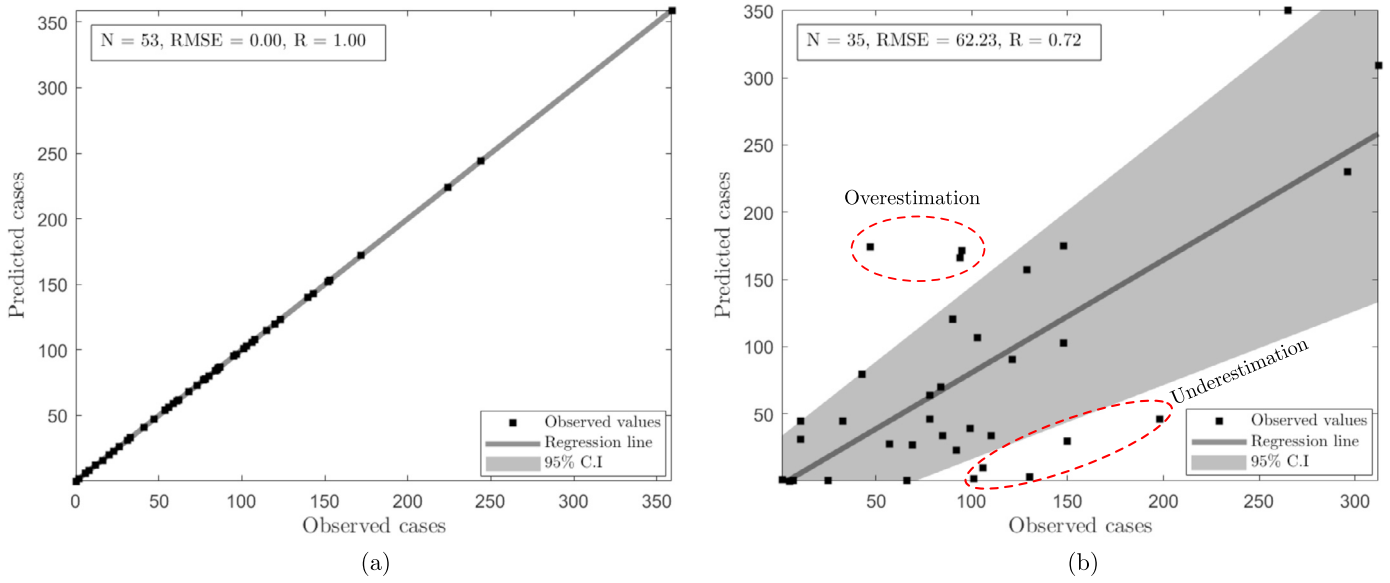

Fig. 9. FF-ANN predicted malaria plotted against the observed cases. (a) for training dataset, and (b) for testing dataset.

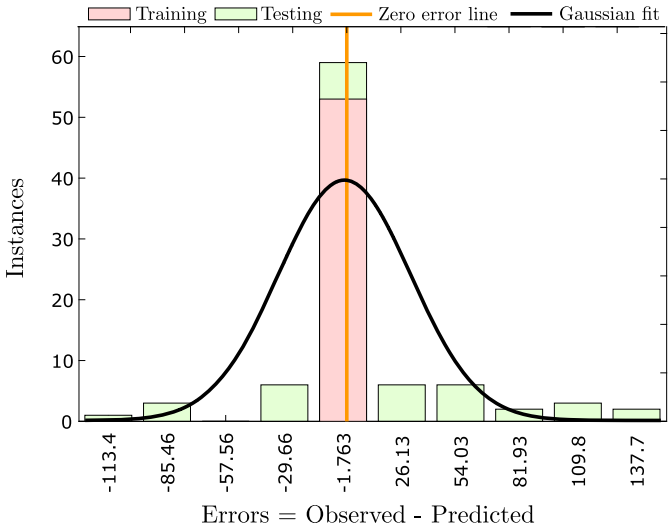

Fig. 10. Error histogram analysis for FF-ANN with 10 bin size. Regions on the left and right of the zero error line (in orange) represent overestimation and underestimation region, respectively.

a peak at zero error line. This indicates for most of the instances output is close to the observed value with occasional underestimation and overestimation.

We performed residual analysis to estimate the appropriateness of the FF-ANN approach (Fig. 11). For a good fit model, the residuals must be randomly scattered without following any deterministic pattern. In other words, the residual must be consistent with the stochastic error. Fig. 11, we observe that although the residuals follow random pattern for both the phases (*i.e.,* training and testing) but a large number of residuals lies outside the testing RMSE line. Hence, the models fail to attain an accuracy equivalent to the training phase over the testing dataset, indicating a case of slight overfitting.

5.3. *Performance of the* P^{2}CA – PSO – ANN

We trained the P^{2}CA – PSO – ANN model by using 60% of the data (N = 53) and evaluated the training accuracy of the proposed approach considering R, RMSE, and bias as the performance metrics. We found that the model performs efficiently on the training data (with R = 0.99, RMSE = 0.01, and bias = -1.31). However, for a fair evaluation, we assessed the performance of the trained model over the remaining 40% of the data (N = 35). We found that the model performs equally well on the unseen data with R = 0.99, RMSE = 1.76, and bias = -1.75. (See Fig. 12.)

We plot histogram to understand the error distribution generated during the prediction of the malaria cases using the proposed P^{2}CA – PSO – ANN model during training and testing. We calculate the error from L1 norms by discarding the absolute part and plotted the stacked histogram of the training (in shades of red) and testing (in shades of green) errors using 10 bin size (Fig. 13). We found that the error ranges from 0.53 (left-most bin) to 10.07 (right-most bin). The total error follows a right skewed distribution. The zero error line lies adjacent to the peak of the distribution. Hence, in most instances, the predicted output is close to the observed values.

We performed the residual analysis of the proposed P^{2}CA – PSO – ANN model (Fig. 14). Unlike in the case of FF-ANN, we found that P^{2}CA – PSO – ANN successfully captures the deterministic part of the response variable. We observed that most of the residuals lie within the testing RMSE line and do not follow any specific pattern (*i.e.,* stochastic in nature) for both training and testing phases indicating that the model is a good fit.

5.4. *Comparison with benchmark algorithms*

For a fair evaluation of the machine learning models, we have compared the results of P^{2}CA – PSO – ANN and FF-ANN with the results of five benchmark algorithms; GRNN, GPR, SVR, Random Forest, and RBNN. We observed that all the algorithms perform differently on the same dataset. The P^{2}CA – PSO – ANN outperforms all the other algorithms in terms of accuracy (Table 2). FF-ANN ranks second in predicting the number of malaria cases. We found the presence of negative bias (*i.e.,* underestimation) in most of the benchmark algorithms. This indicates that all these algorithms underestimate some values except RBNN, which significantly overestimates the malaria cases with a positive bias (58.62).

Although the proposed approach gives promising results, it has some limitations in terms of computational complexity. The use of P^{2}CA – PSO – ANN increases the computational complexity of the proposed approach. For a better comparison, we plot the computational time-complexity graph for all the algorithms (Fig. 15). We observed that P^{2}CA – PSO – ANN exhibits a higher time-computational cost, fol-

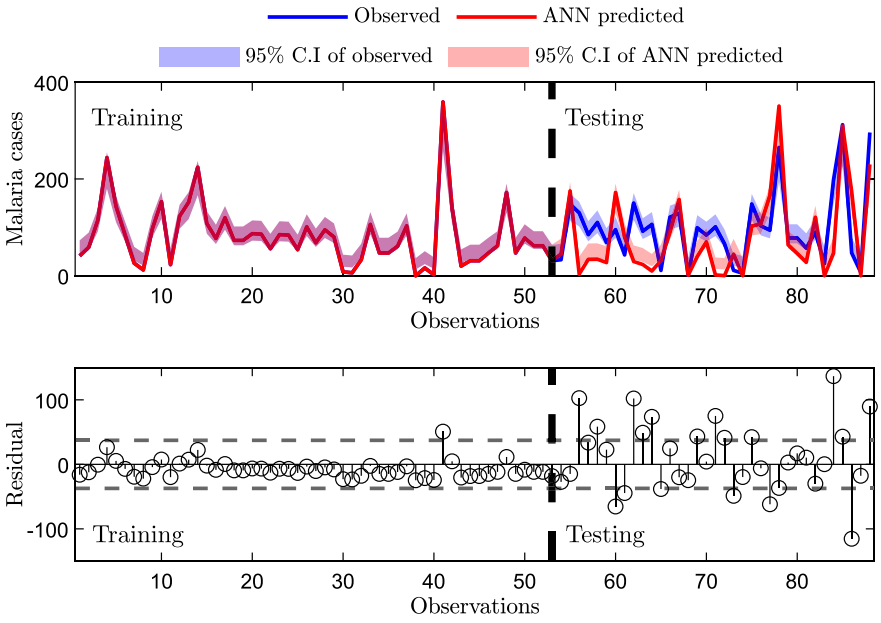

Fig. 11. Time series of the observed vs FF-ANN predicted malaria cases and the corresponding residual plot. Dashed line in the residual shows the testing RMSE.

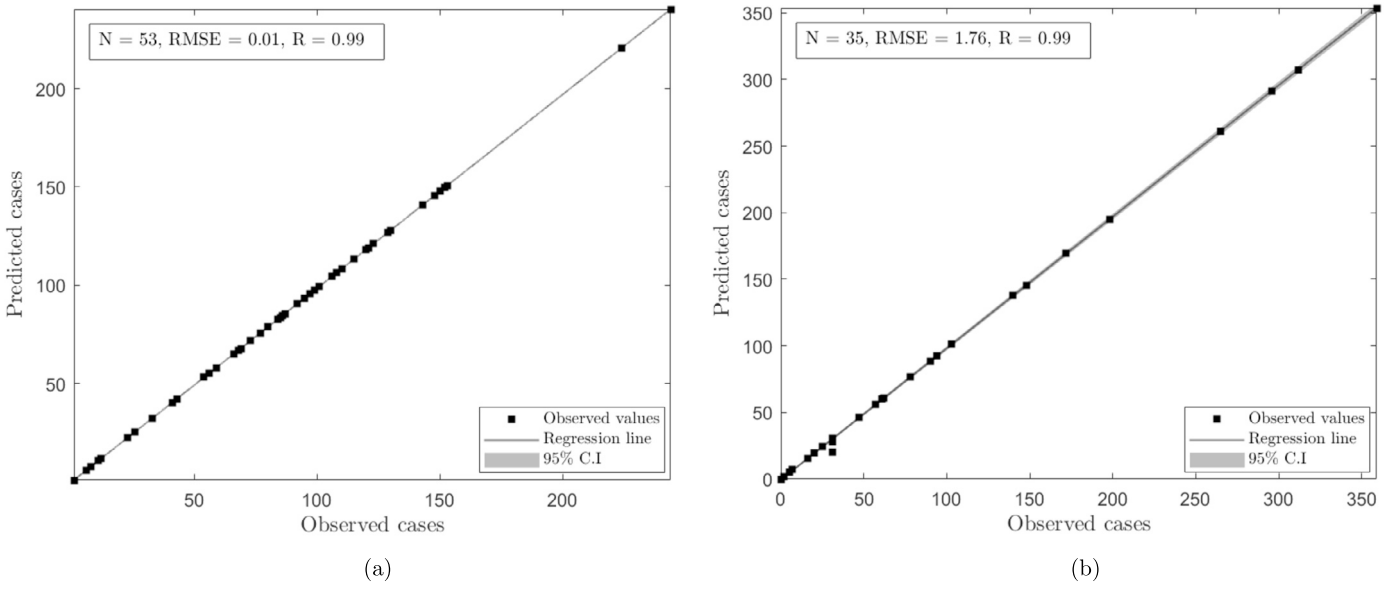

Fig. 12. P^{2}CA – PSO – ANN predicted malaria plotted against the observed cases. (a) for training dataset, and (b) for testing dataset.

Table 2
Comparison of the results with the benchmark algorithms (GRNN, GPR, SVR, Random Forest, and RBNN).

Performance metrics	Methods						
	P^{2}CA-PSO-ANN	FF-ANN	GRNN	GPR	SVR	Random Forest	RBNN
R	0.99	0.72	0.09	0.07	0.44	0.28	0.11
RMSE	1.76	62.23	93	93	92	89.57	92.86
Bias	-1.75	-19.28	-18.48	-55.11	-31.6	-8.97	58.62

lowed by FF-ANN, RBNN, Random Forest, GRNN, and SVR whereas GPR exhibits the least time-complexity. This is primarily because of the computation time the model takes to optimise a large number of internal parameters (*i.e.,* weights and biases) in the case of P^{2}CA – PSO – ANN and FF- ANN. In contrast, those algorithms that have very few free parameters exhibit less computational time, such as GPR, SVR, GRNN, and Random Forest.

5.5. Controlling meteorological variables

We plot the time series of meteorological variables (rainfall, temperature, relative humidity) to assess their control on malaria outbreak (Fig. 16). Rainfall during the monsoon period results in an increase in vector mosquito population. This causes a sharp rise in malaria cases (Fig. 3). During this period relative humidity is relatively high. This input variable has also emerged as the most relevant feature in mapping

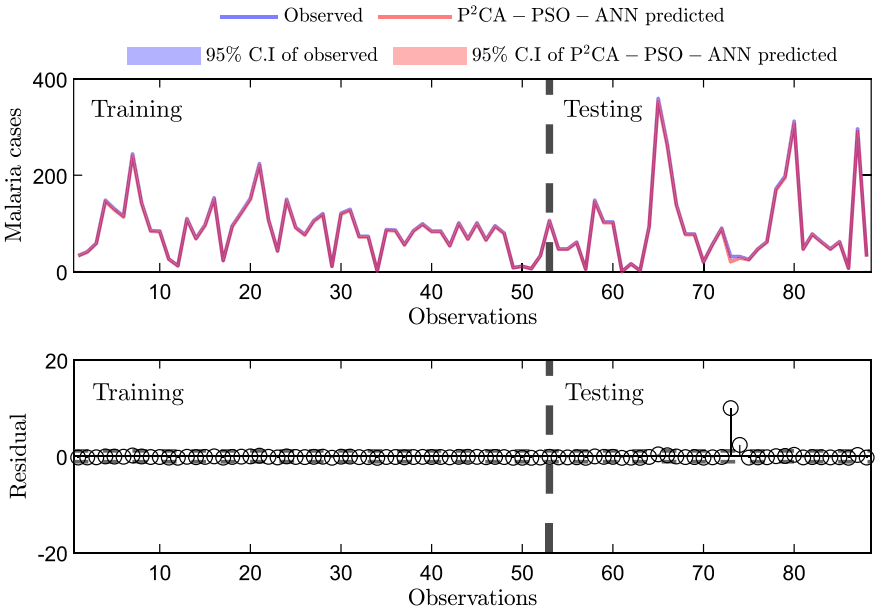

Fig. 14. Time series of the observed vs P²CA – PSO – ANN predicted malaria cases and the corresponding residual plot. Dashed line in the residual shows the testing RMSE.

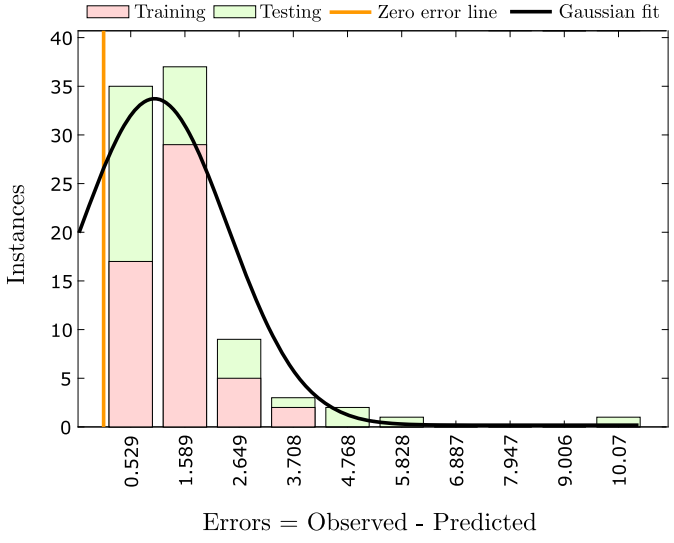

Fig. 13. Error histogram analysis for P²CA – PSO – ANN with 10 bin size. Regions on the left and right of the zero error line (in orange) represent overestimation and underestimation region, respectively.

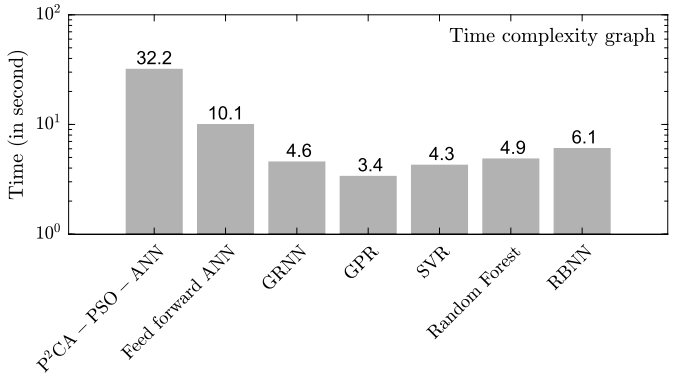

Fig. 15. Time complexity of P²CA – PSO – ANN with the benchmark machine learning algorithms.

malaria cases. Further, the temperature in the study area is relatively high during the summer and decreases greatly at the onset of monsoon. Together with rainfall, high relative humidity, optimal temperature provides favourable conditions for mosquitoes. We have noticed the combined effect of temperature and relative humidity together turns-out to be the second most important relevant feature in mapping malaria case. This suggests temperature and relative humidity are important input predictors that provide favourable initial conditions for the malaria outbreak (Arab et al., 2014).

6. Conclusion

We used meteorological variables to predict malaria outbreak from a hybrid model (*i.e.,* P²CA – PSO – ANN) in three districts of Rajasthan in the western India. Based on the outcomes of this study, we can draw the following conclusions;

- Coupling the probabilistic principal component analysis and particle swarm optimisation with ANN significantly improves the performance.
- The hybrid algorithm accurately predicts the malaria outbreaks. This algorithm is resistance to missing values in the feature set, that increases the robustness of the model.
- Linear data fusion of meteorological variables increases the predictive capability of the machine learning model. The combined effect of relative humidity and temperature shows a high predictive capacity. Relative humidity has emerged as the most important variable in predicting the malaria outbreak.

The outcome of this study can be implemented at the district/state level for early prediction of malaria outbreaks in a region based on the climate forecast. This will help the concerned health departments to take precautionary measures to prevent disease outbreaks. The methodology developed in this study provides encouraging results with limited data. The robustness and prediction curability of our model needs to be evaluated with a long time series of input data. Further, this methodology can be generalised to predict any other types of vector-borne diseases.

In this study, we have only used meteorological variables to predict malaria outbreaks. Another important parameter that greatly controls

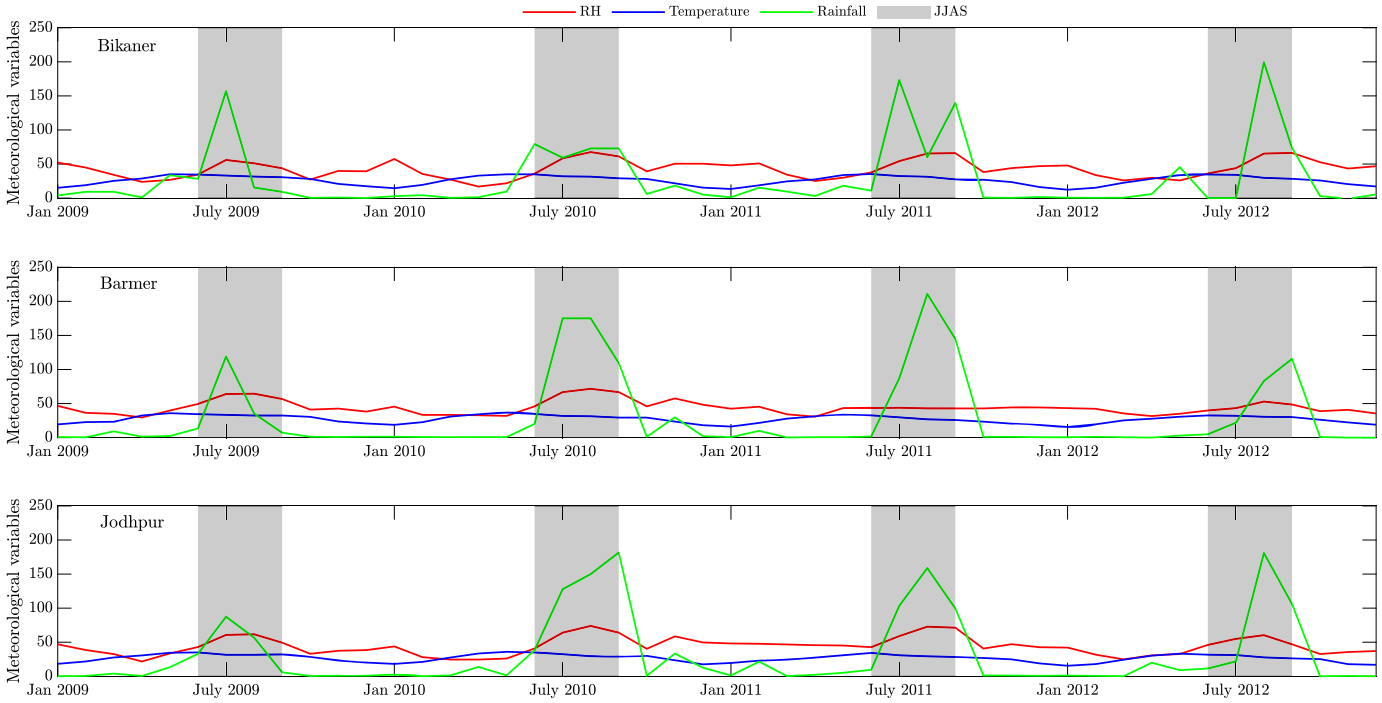

Fig. 16. Time series (form 2009 to 2012) of the meteorological variables (*i.e.,* RH, Temperature, and Rainfall) in three districts (Bikaner, Barmer, and Jodhpur) of Rajasthan in the Western India. The shaded region in grey colour represents the monsoon period (*i.e.*, JJAS; June-July-August-September).

the malaria outbreak is waterlogging during the rainy season (Ding et al., 2014, Podder et al., 2019, Majumdar, 2021). The waterlogging should be included as an input to evaluate its importance in predicting malaria outbreak.

CRediT authorship contribution statement

Abhilash Singh: Conceptualization, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft. **Manish Mehra:** Data curation, Validation, Visualization, Writing – original draft. **Amit Kumar:** Data curation, Visualization. **M Niranjannaik:** Formal analysis, Visualization. **Dev Priya:** Visualization, Writing – review & editing. **Kumar Gaurav:** Investigation, Methodology, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing.

Declaration of competing interest

We have no conflicts of interest to disclose.

Acknowledgements

We would like to acknowledge IISER Bhopal for providing institutional support. AS is thankful to the Department of Science and Technology (DST), Govt. of India for providing DST INSPIRE fellowship (Grant No. DST/INSPIRE Fellowship/[IF180001]).

Appendix A. Performance metrics

We used the following equations to compute the value of R, RMSE, and bias. All these performance metrics are widely used for evaluating the performance of any regression-based machine learning algorithms. The value of R ranges between zero (*i.e.,* worst) to one (*i.e.,* best). The value of RMSE ranges from zero (*i.e.,* best) to infinity (*i.e.,* worst). Bias can be either positive (*i.e.,* overestimation) or negative (*i.e.,* underestimation).

$$R = \sqrt{1 - \frac{Error_{SSE}}{Error_{SST}}} \tag{A.1}$$

$$Error_{SSE} = \sum (y_{obs} - y_{pred})^2 \tag{A.2}$$

$$Error_{SST} = \sum (y_{obs} - \bar{y}_{pred})^2 \tag{A.3}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum (y_{pred} - y_{obs})^2} \tag{A.4}$$

$$Bias = \frac{1}{n}\sum (y_{pred} - y_{obs}) \tag{A.5}$$

where $Error_{SSE}$ is the sum of squares of errors, $Error_{SST}$ is the sum of squares of total, y_{pred} is the predicted value, y_{obs} is the observe value, and n is the number of observations.

References

Akhtar, R., & McMichael, A. (1996). Rainfall and malaria outbreaks in western Rajasthan. *Lancet (London, England)*, *348*, 1457–1458.

Alonso, D., Bouma, M. J., & Pascual, M. (2011). Epidemic malaria and warmer temperatures in recent decades in an East African highland. *Proceedings of the Royal Society B, Biological Sciences*, *278*, 1661–1669.

Alshdaifat, E., Alshdaifat, D., Alsarhan, A., Hussein, F., El-Salhi, S. M. F. S., et al. (2021). The effect of preprocessing techniques, applied to numeric features, on classification algorithms' performance. *Data*, *6*, 11.

Arab, A., Jackson, M. C., & Kongoli, C. (2014). Modelling the effects of weather and climate on malaria distributions in West Africa. *Malaria Journal*, *13*, 1–9.

Asteris, P. G., Roussis, P. C., & Douvika, M. G. (2017). Feed-forward neural network prediction of the mechanical properties of sandcrete materials. *Sensors*, *17*, 1344.

Caminade, C., Kovats, S., Rocklov, J., Tompkins, A. M., Morse, A. P., Colón-González, F. J., Stenlund, H., Martens, P., & Lloyd, S. J. (2014). Impact of climate change on global malaria distribution. *Proceedings of the National Academy of Sciences*, *111*, 3286–3291.

Caminade, C., McIntyre, K. M., & Jones, A. E. (2019). Impact of recent and future climate change on vector-borne diseases. *Annals of the New York Academy of Sciences*, *1436*, 157.

Corte-Valiente, A. D., Castillo-Sequera, J. L., Castillo-Martinez, A., Gómez-Pulido, J. M., & Gutierrez-Martinez, J.-M. (2017). An artificial neural network for analyzing overall uniformity in outdoor lighting systems. *Energies*, *10*, 175.

Di Gennaro, F., Marotta, C., Locantore, P., Pizzol, D., & Putoto, G. (2020). Malaria and Covid-19: Common and different findings. *Tropical Medicine and Infectious Disease*, *5*, 141.

Ding, G., Gao, L., Li, X., Zhou, M., Liu, Q., Ren, H., & Jiang, B. (2014). A mixed method to evaluate burden of malaria due to flooding and waterlogging in Mengcheng County, China: A case study. *PLoS ONE*, *9*, Article e97520.

Dorofki, M., Elshafie, A. H., Jaafar, O., Karim, O. A., & Mastura, S. (2012). Comparison of artificial neural network transfer functions abilities to simulate extreme runoff

data. *International Proceedings of Chemical, Biological and Environmental Engineering*, *33*, 39–44.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 1189–1232.
Garrido-Cardenas, J. A., Cebrián-Carmona, J., González-Cerón, L., Manzano-Agugliaro, F., & Mesa-Valle, C. (2019). Analysis of global research on malaria and plasmodium vivax. *International Journal of Environmental Research and Public Health*, *16*, 1928.
Garrido-Cardenas, J. A., González-Cerón, L., Manzano-Agugliaro, F., & Mesa-Valle, C. (2019). Plasmodium genomics: An approach for learning about and ending human malaria. *Parasitology Research*, *118*, 1–27.
Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, *24*, 44–65.
Hall, E. L., Kruger, R. P., Dwyer, S. J., Hall, D. L., Mclaren, R. W., & Lodwick, G. S. (1971). A survey of preprocessing and feature extraction techniques for radiographic images. *IEEE Transactions on Computers*, *100*, 1032–1044.
Haque, U., Hashizume, M., Glass, G. E., Dewan, A. M., Overgaard, H. J., & Yamamoto, T. (2010). The role of climate variability in the spread of malaria in Bangladeshi highlands. *PLoS ONE*, *5*, Article e14341.
Hulme, M., et al. (1996). *Climate change and Southern Africa: An exploration of some potential impacts and implications in the SADC (Southern African development community) region*. Norwich (United Kingdom) CRU/WWF.
Jetten, T., Martens, W., & Takken, W. (1996). Model simulations to estimate malaria risk under climate change. *Journal of Medical Entomology*, *33*, 361–371.
Jones, A. E., Wort, U. U., Morse, A. P., Hastings, I. M., & Gagnon, A. S. (2007). Climate prediction of El Niño malaria epidemics in North-West Tanzania. *Malaria Journal*, *6*, 1–15.
Karlik, B., & Olgac, A. V. (2011). Performance analysis of various activation functions in generalized MLP architectures of neural networks. *International Journal of Artificial Intelligence and Expert Systems*, *1*, 111–122.
Kelly-Hope, L. A., Hemingway, J., & McKenzie, F. E. (2009). Environmental factors associated with the malaria vectors Anopheles gambiae and Anopheles funestus in Kenya. *Malaria Journal*, *8*, 1–8.
Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks, vol. 4* (pp. 1942–1948). IEEE.
Kim, Y., Ratnam, J., Doi, T., Morioka, Y., Behera, S., Tsuzuki, A., Minakawa, N., Sweijd, N., Kruger, P., Maharaj, R., et al. (2019). Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model. *Scientific Reports*, *9*, 1–10.
Kumar, P., Vatsa, R., Sarthi, P. P., Kumar, M., & Gangare, V. (2020). Modeling an association between malaria cases and climate variables for Keonjhar district of Odisha, India: A Bayesian approach. *Journal of Parasitic Diseases*, 1–13.
Kumar, P., Pisudde, P., & Sarthi, P. P. (2022). Meteorological linkage of malaria cases in the eastern state of India. *The Journal of Climate Change and Health*, *5*, Article 100064.
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*, 764–766.
Lingala, M. A. (2017). Effect of meteorological variables on plasmodium vivax and plasmodium falciparum malaria in outbreak prone districts of Rajasthan, India. *Journal of Infection and Public Health*, *10*, 875–880.
Lingala, M. A. L., Singh, P., Verma, P., & Dhiman, R. C. (2020). Determining the cutoff of rainfall for plasmodium falciparum malaria outbreaks in India. *Journal of Infection and Public Health*, *13*, 1034–1041.
Majumdar, S. (2021). Spatiotemporal pattern and hotspot detection of malaria using spatial analysis and GIS in West Bengal: An approach to medical GIS. In *Healthcare paradigms in the Internet of things ecosystem* (pp. 83–100). Elsevier.
Mathur, K., Harpalani, G., Kalra, N., Murthy, G., & Narasimham, M. (1992). Epidemic of malaria in Barmer district (Thar desert) of Rajasthan during 1990. *Indian Journal of Malariology*, *29*, 1–10.
Modu, B., Polovina, N., Lan, Y., Konur, S., Asyhari, A. T., & Peng, Y. (2017). Towards a predictive analytics-based intelligent malaria outbreak warning system. *Applied Sciences*, *7*, 836.
Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, *22*, Article 100508.
Parihar, R. S., Bal, P. K., Saini, A., Mishra, S. K., & Thapliyal, A. (2022). Potential future malaria transmission in Odisha due to climate change. *Scientific Reports*, *12*, 1–13.
Patz, J. A. (2002). A human disease indicator for the effects of recent global climate change. *Proceedings of the National Academy of Sciences*, *99*, 12506–12508.
Podder, D., Paul, B., Dasgupta, A., Bandyopadhyay, L., Pal, A., Roy, S., et al. (2019). Community perception and risk reduction practices toward malaria and dengue: A mixed-method study in slums of Chetla, Kolkata. *Indian Journal of Public Health*, *63*, 178.
Rocklöv, J., & Dubrow, R. (2020). Climate change: An enduring challenge for vector-borne disease prevention and control. *Nature Immunology*, *21*, 479–483.
Sarkar, S., Gangare, V., Singh, P., & Dhiman, R. C. (2019). Shift in potential malaria transmission areas in India, using the fuzzy-based climate suitability malaria transmission (FCSMT) model under changing climatic conditions. *International Journal of Environmental Research and Public Health*, *16*, 3474.
Singh, A., Kotiyal, V., Sharma, S., Nagar, J., & Lee, C.-C. (2020). A machine learning approach to predict the average localization error with applications to wireless sensor networks. *IEEE Access*, *8*, 208253–208263.
Singh, A., Gaurav, K., Rai, A. K., & Beg, Z. (2021). Machine learning to estimate surface roughness from satellite images. *Remote Sensing*, *13*, 3794.
Singh, A., Nagar, J., Sharma, S., & Kotiyal, V. (2021). A Gaussian process regression approach to predict the k-barrier coverage probability for intrusion detection in wireless sensor networks. *Expert Systems with Applications*, *172*, Article 114603.
Singh, A., Sharma, S., & Singh, J. (2021). Nature-inspired algorithms for wireless sensor networks: A comprehensive survey. *Computer Science Review*, *39*, Article 100342.
Singh, A., Amutha, J., Nagar, J., Sharma, S., & Lee, C.-C. (2022a). Lt-fs-id: Log-transformed feature learning and feature-scaling-based machine learning algorithms to predict the k-barriers for intrusion detection using wireless sensor network. *Sensors*, *22*, 1070.
Singh, A., Amutha, J., Nagar, J., Sharma, S., & Lee, C.-C. (2022b). Automl-id: Automated machine learning model for intrusion detection using wireless sensor network. *Scientific Reports*, *12*, 1–14.
Srimath-Tirumula-Peddinti, R. C. P. K., Neelapu, N. R. R., & Sidagam, N. (2015). Association of climatic variability, vector population and malarial disease in district of Visakhapatnam, India: A modeling and prediction analysis. *PLoS ONE*, *10*, Article e0128377.
Sutherst, R. (1998). Implications of global change and climate variability for vector-borne diseases: Generic approaches to impact assessments. *International Journal for Parasitology*, *28*, 935–945.
Thakur, S., & Dharavath, R. (2019). Artificial neural network based prediction of malaria abundances using big data: A knowledge capturing approach. *Clinical Epidemiology and Global Health*, *7*, 121–126.
Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, *61*, 611–622.
Tolоși, L., & Lengauer, T. (2011). Classification with correlated features: Unreliability of feature ranking and solutions. *Bioinformatics*, *27*, 1986–1994.
Tyagi, B., Chaudhary, R., & Yadav, S. (1995). Epidemic malaria in Thar desert, India. *The Lancet*, *346*, 634–635.
Vogl, T. P., Mangis, J., Rigler, A., Zink, W., & Alkon, D. (1988). Accelerating the convergence of the back-propagation method. *Biological Cybernetics*, *59*, 257–263.
WHO (2020). *World malaria report 2020: 20 years of global progress and challenges*. World Health Organization.
Zhang, X., Zou, D., & Shen, X. (2018). A novel simple particle swarm optimization algorithm for global optimization. *Mathematics*, *6*, 287.